Discriminative Cellets Discovery for Fine-Grained Image Categories Retrieval

Luming Zhang¹, Yi Yang², Roger Zimmermann¹
¹School of Computing, National University of Singapore
²School of Information Technology and Electrical Engineering, University of Queensland

ABSTRACT

Fine-grained image categories recognition is a challenging task aiming at distinguishing objects belonging to the same basic-level category, such as leaf or mushroom. It is a useful technique that can be applied for species recognition, face verification, and etc. Most of the existing methods have difficulties to automatically detect discriminative object components. In this paper, we propose a new fine-grained image categorization model that can be deemed as an improved version spatial pyramid matching (SPM). Instead of the conventional SPM that enumeratively conducts cell-to-cell matching between images, the proposed model combines multiple cells into cellets that are highly responsive to object fine-grained categories. In particular, we describe object components by cellets that connect spatially adjacent cells from the same pyramid level. Straightforwardly, image categorization can be casted as the matching between cellets extracted from pairwise images. Toward an effective matching process, a hierarchical sparse coding algorithm is derived that represents each cellet by a linear combination of the basis cellets. Further, a linear discriminant analysis (LDA)-like scheme is employed to select the cellets with high discrimination. On the basis of the feature vector built from the selected cellets, fine-grained image categorization is conducted by training a linear SVM. Experimental results on the Caltech-UCSD birds, the Leeds butterflies, and the COSMIC insects data sets demonstrate our model outperforms the state-of-the-art. Besides, the visualized cellets show discriminative object parts are localized accurately.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software; I.4.8 [Image Processing and Computer Vision]: Scene Analysis and Sensor Fusion

Keywords
Fine-grained, Categories retrieval, Sparse coding, Cellets,

Spatial pyramid

1. INTRODUCTION

In the past a few years, many object recognition models have been proposed for multimedia retrieval and analysis [33, 34, 3]. Most of them focus on discriminative learning for detecting and localizing instances belonging to different basic-level categories (e.g., car, cow, and human). Recently, motivated by the application in areas such as agriculture, medicine, and forestry, fine-grained domain recognition has become a hot research topic [4, 5]. For example, researchers are designing image retrieval models to recognize pests found on and around some farm crops such that their species can be monitored. Thereby, suitable chemicals can be taken to mitigate these pests. However, it is still challenging to deal with fine-grained categorization successfully due to two reasons: 1) the difficulty to automatically discover the arbitrary-shaped object components as the examples shown in Fig. 1; and 2) the dynamic backgrounds, occlusions, and variations in lighting conditions, leading to obstacles to learn a robust fine-grained categorization model.

To solve the above problems, we propose a new fine-grained image categorization model that integrates the spatially adjacent cells from a pyramid into discriminative ones. These spatially adjacent cells are called cellets and describe the discriminative object components in a coarse-to-fine manner, as shown in Figure 2. An overview of the proposed framework is presented as follows. By dividing each image into multi-level cells hierarchically, we construct cellets by connecting spatially adjacent cells to describe object components with different scales. To calculate the similarity...
between images, it is straightforward to match all their corresponding cellets. However, the enumeratively cellet-to-cellet matching is computationally intractable. Toward an efficient matching process, a hierarchical sparse coding algorithm is proposed. The first layer seeks the basis cellets with the same structure; while the second layer finds the basis for cellets with different structures. Afterward, each cellet can be represented by a linear combination of the basis cellets. To accelerate the cellet matching process, we derive the discrimination of a cellet for selecting cellets highly descriptive to the fine-grained categories. Finally, we represent each image by a set of discriminative cellets, which are further converted into an image-level feature vector. These feature vectors are stored for training a linear SVM for fine-grained image categorization.

The main contributions of this work can be summarized as follows: 1) we mine image descriptors highly descriptive to fine-grained categories, based on which an image recognition model is developed, 2) a hierarchical sparse coding algorithm for cellets quantization is introduced, and 3) the first experimental result on a fine-grained data set containing over 5,000 real-world insects is released.

2. RELATED WORK

The proposed fine-grained categorization algorithm is closely related to two research topics in multimedia analysis and pattern recognition: the spatial pyramid matching (SPM) architecture and the fine-grained image categorization.

Toward a general categorization model, SPM is developed that reflects the rough image geometric characteristics. Some researchers [27] have pointed out that the k-means-based codebook in SPM is less effective. Thus, sparse coding to represent each by a linear combination of the basis cellets. In the encoding stage, the learned codebook is applied to a new set of local descriptors $X'$ to obtain the encoded local descriptors, i.e.,

$$\min_{y_{m}, \mathbf{B}} \sum_{m} ||x_{m} - y_{m}\mathbf{B}||_2^2 + \lambda_1 ||y_{m}||_1, \quad (1)$$

where $\mathbf{B} \in \mathbb{R}^{R \times D}$ is a codebook with $R$ basis vectors. Normally, codebook $\mathbf{B}$ is an over-complete basis set, i.e., $R > D$.

In the encoding stage, the learned codebook $\mathbf{B}$ is applied to a new set of local descriptors $X'$ to obtain the encoded local descriptors, i.e.,

$$\min_{Y} \sum_{m} ||x_{m} - y_{m}\mathbf{B}||_2^2 + \lambda_1 ||y_{m}||_1, \quad (2)$$

where $Y = [y_1, y_2, \ldots, y_M] \in \mathbb{R}^{M \times R}$ is the encoded local descriptors of local descriptor from $X'$.

To incorporate the spatial information into local descriptors, a spatial pyramid is constructed. In detail, based on the coordinates of the local descriptors, we divide the local descriptors into several cells $\{Y_{ij}\}$, where the script denotes the $ij$-th cell from the $l$-th layer spatial pyramid. Formally, the location of cell $Y_{ij}$ is defined as: $c(Y_{ij}) = (l, i, j)$.

Different cells may contain different numbers of local descriptors. Toward a fixed length of feature vector for each cell, the max pooling scheme [6] is applied on each cell:

$$u'_{ij} = \xi(Y'_{ij}), \quad (3)$$

where $\xi$ denotes the maximum element on each row of $Y'_{ij}$; and $u'_{ij}$ denotes a $R$-dimensional column feature vector.

As we discussed above, the SPM architecture cannot fulfill fine-grained categorization. This is because the cell-to-cell matching includes visual features non-discriminative to object components. To tackle this problem, we propose cellets for spatial pyramid matching. A cellet denotes a set...
of spatially adjacent cells $U = \{u_i\}$ associated with their structure, which can be defined as follows:

$$z = [\psi(U), \phi(U)]^T.$$  

(4)

The first term $\psi(U) = \cup_{u \in U}[u^T]$ denotes a set of spatially adjacent cells, wherein $\cup[.]$ is a vector concatenation operator. As shown on the left of Figure 3, cellets with the same number of cells may have different structures. These structures are discriminative cues that can contribute to the fine-grained categorization. Thus, we use the second term $\phi(U) = \text{vec}(\phi(U))$ to represent the structure of a cellet. Here, vec(·) is a row-wise vector stacking operator, and $\varphi$ is the binary spatial relations between cells in $U$:

$$\varphi(i, j) = \begin{cases} k & \text{if } \frac{k \pi}{K} \leq \theta(u_i, u_j) \leq \frac{(k + 1) \pi}{K} \\ 0 & \text{otherwise} \end{cases},$$

(5)

where $\theta(u_i, u_j)$ denotes the angle between the spatially adjacent cells $u_i$ and $u_j$, and $K$ determines the minimum angle that can be discriminated. We can set $K$ to 4 or 8 depending on different data sets, as shown on the right of Figure 3.

Based on the location of a cell, we define the location of a cellet $z$ by combining the locations of all the cells in $z$, i.e.,

$$\kappa(z) = \cup_{u \in U}[\kappa(u)].$$

(6)

In summary, the flowchart of generating cellets from an image is presented in Figure 4.

4. HIERARCHICAL CELLETS ENCODING

Conventional SPM methods cast image categorization as the cellet-to-cellet matching between images. However, there are numerous cellets in a spatial pyramid. Denote $|z|$ as the number of cells that construct cellet $z$, the number of cellets from the $l$-th layer of spatial pyramid is no less than:

$$H = 4^{l-1} \cdot 4 \cdot 6 \cdots \cdot |2| = 4^{l-1} |z|2^{l-1}.$$  

(7)

Thus, given a moderate cellet size, it is intractable to compare pairwise images by enumeratively cellet-to-cellet matching. For fine-grained recognition task, only cellets discriminative to object details should be preserved.

4.1 Hierarchical Sparse Coding of Cellets

To measure the discrimination of a cellet, first we need to measure the similarity between cellets. Because the term $\phi(U)$ in (4) is un-quantized, it is necessary for us to learn a codebook from the training cellets. The codebook allows to represent a cellet by a linear combination of the basis cellets. However, due to the large number of cellets, it is intractable to learn the codebook by employing the training cellets once-for-all, as the standard sparse coding [13].

To solve this problem, a hierarchical sparse coding is proposed that decomposes the encoding task on a large number of cellets into a set of sub-procedures. In the first layer, the basis cellets with the same structure are derived. Particularly, given $N$ training images, we collect cellets with the same structure:

$$Z = [z_1, z_2, \cdots, z_N],$$

(8)

where $z_i$ is the cellet from the $i$-th training image.

Then, we use sparse coding to find a set of basis cellets:

$$\min_{\mathbf{D}_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} ||z_i - \mathbf{D}_1 \alpha_i||_2^2 + \lambda_2 ||\alpha_i||_1 \right\},$$

(9)

where $\mathbf{D}_1 \in \mathbb{R}^{K \times T}$ is a codebook learned from $Z$ and each column of $\mathbf{D}_1$ represents a basis cellet. After the first layer sparse coding, each cellet can be represented by a linear combination of basis cellets with the same structure.

Because the number of basis cellets in $\mathbf{D}_1$ is still large, the second layer sparse coding is proposed to find basis cellets with different structures. Following (7), we obtain $H$ codebooks from the first layer sparse coding:

$$\mathbf{D}_1 = [\mathbf{D}_1^1, \mathbf{D}_1^2, \cdots, \mathbf{D}_1^K].$$

(10)

Thereafter, the second layer sparse coding learns a codebook $\mathbf{D}_2$ from $\mathbf{D}_1$, i.e.,

$$\min_{\mathbf{D}_2} \left\{ \frac{1}{HT} \sum_{i=1}^{HT} ||d_i - \mathbf{D}_2 \beta_i||_2^2 + \lambda_3 ||\beta_i||_1 \right\},$$

(11)

where $d_i$ denotes the $i$-th column of matrix $\mathbf{D}_1$. After the second layer sparse coding, given a new cellet $z_{\text{test}}$, we represent it by a linear combination of the basis cellets:

$$\min_{\beta} \left\{ ||z_{\text{test}} - \mathbf{D}_2 \gamma||_2^2 + \lambda_3 ||\gamma||_1 \right\},$$

(12)

where $\gamma(z_{\text{test}})$ is the sparse representation of cellet $z_{\text{test}}$.

The algorithm of our proposed hierarchical sparse coding of cellets is summarized below.

**Algorithm 1 Hierarchical Sparse Coding of Cellets**

**input**: A set of training images $\{I_1, I_2, \cdots, I_N\}$, the size of cellet $|z|$, a test cellet $z_{\text{test}}$;

**output**: the sparse representation of cellet $z_{\text{test}}$;

1. Extract SIFT descriptors for each image; use sparse coding to encode them based on (1) and (2);
2. Construct spatial pyramid for each image and obtain cellets with size $|z|$ according to (4);
3. Partition the cellets according to their structure for cellets with the $i$-th structure do
   - Compute the first layer codebook $\mathbf{D}_1^i$ from (9);
   - end for
4. Obtain the sparse representation of $z_{\text{test}}$.

4.2 Selecting Discriminative Cellets

To select cellets descriptive to the fine-grained categories, we need to measure the discrimination of a cellet. Inspired by the Fisher’s linear discriminant analysis [1], the measure of a cellet’s discrimination can be defined as:

$$d(z) = \sum_{z'} |\gamma(z) - \gamma(z')| \cdot |\sigma(z, z')|,$$

(13)

where $\sigma(z, z')$ and $\sigma'(z, z')$ are functions indicating whether cellets $z$ and $z'$ belonging to the same category, i.e., if $z$ and $z'$ belong to different categories, then $\sigma(z, z') = 1$ and $\sigma'(z, z') = 0$; otherwise $\sigma(z, z') = 0$ and $\sigma'(z, z') = 1$. A
larger \(d(z)\) reflects a higher discrimination of cellet \(z\).

In order to compute (13), we have to employ all the training cellets to compare with cellet \(z\), which is highly time-consuming. To accelerate it, following the SPM architecture, only cellets \(z'\) with the same location of cellets \(z\) are compared. Therefore, (13) can be reformulated as:

\[
d'(z) = \frac{\sum_{z' : \lambda(z') = \lambda(z)} |\gamma(z) - \gamma(z')| \cdot \sigma(z, z')}{\sum_{z' : \lambda(z') = \lambda(z)} |\gamma(z) - \gamma(z')| \cdot \sigma(z, z')} \tag{14}
\]

It is worth emphasizing another advantage of replacing (13) by (14). In the training stage, the locations of the discriminative cellets are learned implicitly. These learned locations can guide the extraction of the discriminative cellets from the test images. That is to say, there is no need to compute the discrimination of cellets in the test stage, which accelerates the test stage remarkably.

5. THE SVM CLASSIFICATION MODEL

Based on the discrimination measure, for each image, we select fixed number of cellets closely related to its category. These cellets capture the object categorical cues in a coarse-to-fine manner, which can be converted into a vector as:

\[
\gamma(Z^d) = \{\gamma(z) \mid \exists z \in Z : \gamma(z) \}\tag{15}
\]

Denoting \(C\) as the number of categories, an SVM-based fine-grained classification is carried out by using a pairwise scheme. And \(C(C - 1)/2\) classifiers are obtained, each of which is trained by images from two different classes. For images from the \(p\)-th class and those from the \(q\)-th class, we construct a linear SVM classifier as:

\[
\min_{\alpha \in \mathbb{R}^{N_{pq}}} \left\{||\alpha||^2 + C \sum_{i=1}^{N_{pq}} l(\alpha, (\gamma(Z^d_i), c_i)) \right\}, \tag{16}
\]

where \(\gamma(Z^d_i)\) is the feature vector from the \(i\)-th training image; \(c_i \in \{1, 2, \cdots, C\}\) is the category to the \(i\)-th training image; \(\alpha\) determines the hyper-plane to separate images in the \(p\)-th category from those in the \(q\)-th category; and \(N_{pq}\) is the number of training images either from the \(p\)-th category or from \(q\)-th category.

Given a feature vector \(\gamma(Z^d_{test})\) from a test image, its category \((p \text{ or } q)\) is predicted by:

\[
\text{sgn}(\sum_{i=1}^{N_{pq}} \alpha_i \gamma(Z^d_{test}) + b). \tag{17}
\]

During test, \(C(C - 1)/2\) times classification will be conducted and the voting rule is utilized for the final decision. That is, each binary classifier can be deemed as a voting process, and \(\gamma(Z^d_{test})\) is assigned to the class with the maximum number of votes.

To summarize the discussion above, we present the pipeline of our proposed model in Algorithm 2.

5.1 Time Complexity Analysis

The time consumption of our approach is as follows: In the training stage, Step 1 and Step 3 respectively contain one sparse coding; the time consumption of the linear SVM training in Step 4 is \(\mathcal{O}(N)\); and Step 2 contains \(H\) times sparse coding. In our experiment, \(H\) ranges from 100 to 5000, reflecting that the training time consumption is largely determined by the efficiency of sparse coding. Practically, there are many off-the-shelf efficient sparse coding solvers, such as that proposed by Lee et al. [13]. Thus, the training time consumption is acceptable. Different from the training phase, the test stage can be carried out rapidly. This is because the time complexity of Step 1 and Step 2 are \(\mathcal{O}(|Z^d|)\) and \(\mathcal{O}(1)\) respectively, where \(|Z^d|\) denotes the number of the selected discriminative cellets.

6. EXPERIMENTAL RESULTS

This section justifies the effectiveness of the proposed algorithm based on three experiments. The first experiment compares the proposed cellet-encoded SPM with the other SPM variants and fine-grained categorization models. The second experiment test the influence of different parameter settings. Last but not least, we visualize the discovered cellets, which illustrates the high accuracy of our approach.

Toward a comprehensive evaluation, we compiled a new dataset termed COSMIC insects, containing insects from 15 categories. The 15 categories are listed in Table 1. In addition, we also experiment on the Caltech-UCSD birds [31] and the Leed butterflies [32]. The experiments are carried out on a computer equipped with an Intel Xeon X5482 CPU and 8GB RAM. All the comparative algorithms are implemented on the Matlab 2011 platform.

Algorithm 2 Cellet-Encoded Spatial Pyramid for Fine-grained Image Categorization

//training stage
input: A set of training images \(\{I_1, I_2, \cdots, I_N\}\) labeled by the fine-grained image categories;
output: Discriminative cellets, a multi-class linear SVM;
1. For image \(I_i\), obtain its cellets \(\{z_1^i, z_2^i, \cdots, z_{N^d_i}^i\}\); Represent each cellet by a set of basis cellets by Alg. 1;
2. For each cellet, compute its discrimination from (14); select the discriminative ones; and represent training image \(I_i\) by a feature vector \(\gamma(Z^d_i)\) according to (15);
3. Train a multi-class linear SVM based on (16).
//test stage
input: A test image \(I_{test}\); output: The category of \(I_{test}\);
1. Obtain the discriminative cellets \(Z^d_{test}\) of image \(I_{test}\) based on the location of the discriminative cellets;
2. Compute the feature vector \(\gamma(Z^d_{test})\); classify it using the trained linear SVM classifier.

Figure 4: The flowchart of cellets generation (\(X \rightarrow Y\): sparse coding of local descriptors; \(Y \rightarrow U\): max pooling of encoded local descriptors in each cell; \(U \rightarrow Z\): cellets by extracting spatially adjacent cells).
Table 1: Statistics of the COSMIC insects data set

<table>
<thead>
<tr>
<th>Insect</th>
<th>Aphids</th>
<th>Armyworm</th>
<th>Boomwax</th>
<th>Colorado</th>
<th>DFM</th>
<th>FleaBeetle</th>
<th>Jassids</th>
<th>Leaf Roller</th>
</tr>
</thead>
<tbody>
<tr>
<td>340</td>
<td>653</td>
<td>444</td>
<td>446</td>
<td>437</td>
<td>237</td>
<td>136</td>
<td>219</td>
<td></td>
</tr>
<tr>
<td>MealYbugs</td>
<td>575</td>
<td>114</td>
<td>202</td>
<td>417</td>
<td>222</td>
<td>302</td>
<td>445</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison of categorization accuracies on the three data sets

<table>
<thead>
<tr>
<th>Method</th>
<th>Caltech-UCSD</th>
<th>Leeds</th>
<th>COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% train</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPM</td>
<td>85.3%</td>
<td>51.6%</td>
<td>42.2%</td>
</tr>
<tr>
<td>SC-SPM</td>
<td>88.9%</td>
<td>32.4%</td>
<td>44.4%</td>
</tr>
<tr>
<td>LLC-SPM</td>
<td>86.7%</td>
<td>31.9%</td>
<td>45.2%</td>
</tr>
<tr>
<td>Our</td>
<td>41.3%</td>
<td>36.4%</td>
<td>45.7%</td>
</tr>
<tr>
<td>50% train</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPM</td>
<td>40.1%</td>
<td>36.4%</td>
<td>63.2%</td>
</tr>
<tr>
<td>SC-SPM</td>
<td>44.3%</td>
<td>39.1%</td>
<td>63.8%</td>
</tr>
<tr>
<td>LLC-SPM</td>
<td>44.1%</td>
<td>37.7%</td>
<td>64.9%</td>
</tr>
<tr>
<td>Our</td>
<td>47.8%</td>
<td>41.3%</td>
<td>65.4%</td>
</tr>
</tbody>
</table>

6.1 Comparison with the Existing Methods

In this subsection, we first compare our method with the conventional SPM [2] and its two variants: SC-SPM [6], LLC-SPM [27]. The Matlab codes of all the three compared methods are publicly available\(^1\). The parameter settings of the compared methods are as follows. For SPM, SC-SPM and LLC-SPM, we construct a three level spatial pyramid; then we extract over one million SIFT descriptors from 16 × 16 patches computed over a grid with spacing of 8 pixels from all training images. Finally, a codebook with size 256 is generated by k-means [12] clustering on these SIFT descriptors. For each of the three data sets, 30% and 50% images are used respectively for training, while the rest are for testing. We report the categorization accuracy in Table 2. As can be seen, the proposed method outperforms SPM and its two variants, since the discovered cellets are more descriptive to object parts than the conventional cells.

In addition, we compare our method with four existing fine-grained categorization models that are proposed by Yao et al. [21], Berg et al. [23], Duan et al. [25], and Angelova et al. [26] respectively. We implement all the four algorithms because their codes are unavailable. Different proportion of training images are used by selecting 10% to 90% training images. As shown in Figure 5, the proposed approach beats all the compared methods. Further, the per-category accuracy on the Caltech-UCSD birds is presented in Figure 6. We compare our approach with Duan et al.’s approach, which is the second best in Figure 5. As can be seen, for most categories, the proposed method outperforms Duan et al.’s approach significantly.

6.2 Effects of Different Parameter Settings

In this subsection, we study the influence of different parameter settings on the three aforementioned data sets. Particularly, we first set the default values of the parameters as detailed in Table 3. Then, we tune one of the parameters and report the categorization accuracy correspondingly. For convenience, the codebook size and the regularization parameter of the dual sparse codings are set to be equal.

1) We report the performance of our approach with different codebook sizes in Table 4. As seen, by increasing the codebook size from 128 to 256, 512 and 1024, the recognition accuracy improves dramatically. But the improvement becomes smaller when the codebook size is larger than 512.

2) Then we report the recognition accuracy under different regularization parameter of sparse coding. More specifically, we choose the regularization parameter from \([0, 10^{-4}, 10^{-5}, 10^{-2}, 10^{-1}]\) and report the categorization accuracy in Table 5. As can be seen, the highest categorization is always achieved when the regularization parameter is \(10^{-3}\) or \(10^{-2}\).

3) Next, we report recognition accuracy with different size of cellets (under 30% training samples on each data set). As seen from Figure 7, when the cellet size is tuned from 1 to 10, the categorization accuracy increases moderately and steadily, but the time consumption increased sharply. This observation demonstrates that in practice, we should choose an optimal cellet size.

4) Then, we present the categorization accuracy with different

Table 3: The default parameter settings

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Caltech-UCSD</th>
<th>Leeds</th>
<th>COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codebook size</td>
<td>256</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>Regular. para.</td>
<td>(10^{-1})</td>
<td>(10^{-1})</td>
<td>(10^{-1})</td>
</tr>
<tr>
<td>Cellet size</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Value of K</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Pyramid level</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4: Comparison of recognition accuracy on the Caltech-UCSD, the Leeds and our own data set

<table>
<thead>
<tr>
<th>Codebook size</th>
<th>Caltech-UCSD</th>
<th>Leeds</th>
<th>COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% train</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td></td>
<td>41.2%</td>
<td>41.9%</td>
<td>42.4%</td>
</tr>
<tr>
<td></td>
<td>45.8%</td>
<td>36.4%</td>
<td>37.1%</td>
</tr>
<tr>
<td></td>
<td>44.5%</td>
<td>39.3%</td>
<td>40.3%</td>
</tr>
<tr>
<td>50% train</td>
<td>47.1%</td>
<td>47.8%</td>
<td>48.4%</td>
</tr>
<tr>
<td></td>
<td>45.5%</td>
<td>41.3%</td>
<td>41.8%</td>
</tr>
<tr>
<td></td>
<td>54.5%</td>
<td>62.1%</td>
<td>54.5%</td>
</tr>
</tbody>
</table>
concepts. Inspired by this concept, the training stage is efficient, while the testing time consumption is very small. Remarkably, the cloud computing technique has become a hot research area. Different models have been proposed to perform fine-grained categorization. Experimental results thoroughly demonstrate the advantage of the proposed model.

5) Last but not least, we report the categorization accuracy with different values of $K$, the minimum angle that can be discriminated. As shown from Table 6, the categorization accuracy increases slightly from $K = 4$ to $K = 8$ on the Leeds and the Caltech-UCSD data sets.

6.3 Visualization of Discriminative Cellets

A unique property of our approach is the “transparency” of the fine-grained visual features extraction process. As shown in Figure 8, we visualize the most discriminative cellets from the Caltech-UCSD birds, the COSMIC insects, and the Leeds butterflies data sets. Due to space limitation, only cellets from the last layer spatial pyramid are shown, reflecting the most detailed object discriminative components. As can be seen, the discriminative cellets from different categories have significantly different appearances and structures, which demonstrates the effectiveness of our proposed model. Further, the selected cellets contain little background information, reflecting that our model is robust to the dynamic backgrounds.

7. CONCLUSIONS

This paper proposes a novel fine-grained image categorization framework. By introducing cellet to represent the spatial layout of an image, we cast fine-grained categorization as the matching between cellets from pairwise images. Then, we develop a hierarchical sparse coding algorithm that represents each cellet by a linear combination of the basis cellets. Finally, the discrimination of cellet is derived for selecting a few discriminative cellets for fine-grained categorization. Experimental results thoroughly demonstrate the advantage of the proposed model.

The training time of the proposed method is moderately large, while the testing time consumption is very small. Recently, the cloud computing technique has become a hot research area. Inspired by this concept, the training stage of the proposed method can be computed in a distributed way, e.g., based on a workstation. And the test stage can be calculated based on a the cell phone. Therefore, the training and testing can both be carried out efficiently.

8. ACKNOWLEDGMENT

Table 5: Performance under different regularization parameter

<table>
<thead>
<tr>
<th>Regu. Para.</th>
<th>$0$</th>
<th>$10^{-4}$</th>
<th>$10^{-3}$</th>
<th>$10^{-2}$</th>
<th>$10^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% train</td>
<td>Caltech-UCSD</td>
<td>34.5%</td>
<td>38.7%</td>
<td>41.9%</td>
<td>41.1%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>27.6%</td>
<td>34.6%</td>
<td>36.4%</td>
<td>35.4%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>35.6%</td>
<td>38.7%</td>
<td>40.3%</td>
<td>37.2%</td>
</tr>
<tr>
<td>50% train</td>
<td>Caltech-UCSD</td>
<td>37.6%</td>
<td>42.4%</td>
<td>45.1%</td>
<td>47.8%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>32.8%</td>
<td>39.2%</td>
<td>38.4%</td>
<td>41.3%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>55.7%</td>
<td>58.7%</td>
<td>59.3%</td>
<td>61.2%</td>
</tr>
</tbody>
</table>

Table 6: Performance under different regularization parameter

<table>
<thead>
<tr>
<th>$K$</th>
<th>$4$</th>
<th>$8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% train</td>
<td>Caltech-UCSD</td>
<td>41.9%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>36.4%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>39.7%</td>
</tr>
<tr>
<td>50% train</td>
<td>Caltech-UCSD</td>
<td>47.8%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>41.3%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>57.6%</td>
</tr>
</tbody>
</table>

Table 7: Performance under different pyramid level

<table>
<thead>
<tr>
<th>$L$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>30% train</td>
<td>Caltech-UCSD</td>
<td>26.4%</td>
<td>36.9%</td>
<td>41.9%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>21.2%</td>
<td>34.1%</td>
<td>36.4%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>28.5%</td>
<td>32.2%</td>
<td>37.2%</td>
</tr>
<tr>
<td>50% train</td>
<td>Caltech-UCSD</td>
<td>30.2%</td>
<td>35.1%</td>
<td>41.9%</td>
</tr>
<tr>
<td></td>
<td>Leeds</td>
<td>26.2%</td>
<td>30.9%</td>
<td>39.4%</td>
</tr>
<tr>
<td></td>
<td>COSMIC</td>
<td>34.4%</td>
<td>34.3%</td>
<td>37.5%</td>
</tr>
</tbody>
</table>
9. REFERENCES


[8] Harchaoui Z. and Bach, F. Image Classification with Segmentation Graph Kernels. in Proc. of CVPR, pages:
Figure 7: Top: categorization accuracy v.s. size of cellets; bottom: Logarithm of time consumption v.s. size of cellets.

1–8, 2007.


